**94-881: Managing Analytics Projects (Fall 2023)**

**HW #3: Choices around Analytics and Visualization**

**Group: Aditi Gupta, Dollaya Hirunyasiri, Greta Luo, Jorge Palacio**

**Topic: Reducing Unplanned Hospital Readmissions**

**1. What analytics problem type matches this decision? Problem types should be at the level of classification, regression, clustering etc., not too high level (e.g., analytics) or specific analytics technique (e.g. random forest)**

Our goal is to determine the most influential factors predicting hospital readmission for diabetic patients. The appropriate analytics approach for this goal is **classification**, as we aim to estimate the probability that a diabetic patient will be readmitted to the hospital within 30 days.

Classification is the best choice because of several reasons.

* **Binary Outcome:** The problem is inherently binary in nature: a patient is either readmitted within 30 days (class 1) or not (class 0). Classification algorithms are specifically designed for predicting distinct classes.
* **Probability Estimates**: Classification models like logistic regression provide not just a binary outcome but also the probability of an event occurring. This can help hospitals in risk stratification. For example, patients with a 90% probability of readmission might need different interventions compared to those with a 60% probability.
* **Performance Metrics**: Classification tasks offer a variety of performance metrics such as accuracy, precision, recall, F1-score, and AUC-ROC, among others. This allows healthcare institutions to choose a metric that aligns best with their goals (e.g., maximizing true positives while minimizing false negatives).
* **Imbalance Handling**: In our situation, the number of patients readmitted within 30 days is much lower than those who aren't. Many classification algorithms have techniques to handle class imbalances, such as oversampling, undersampling, or using anomaly detection methods.

**2. What analytics techniques can be considered for this problem type? Full answers should list more than one technique and be specific. (not classification – too high-level but algorithms like logistic regression and random forests)**

Our recommended technique for predicting hospital readmission for diabetic patients within 30 days is decision trees. In answers to questions 2 and 3, we’ll show the suitable techniques for classification problems, the constraints of the readmission problem, and how these techniques fulfill those restrictions.

Suitable techniques for classification problems:

**Decision Tree:**

Description:

• A decision tree uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes.

• Algorithms for constructing decision trees usually work top-down, by choosing a variable

(“Race”, “Gender”, “Consumption of insulin”) at each step that best splits the set of items with regard to the target variable (“Readmitted within 30 days”).

• Best split is the one that results in the lowest probability that the results could have been generated by chance.

Advantages:

• Easy to understand, interpret and visualize.

• Handles Mixed Data: It can handle a mix of categorical and numerical features, making it suitable for healthcare datasets that often contain diverse types of information.

• The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.6

Use Cases:

• Decision turn is widely used in healthcare for tasks like disease prediction, readmission risk assessment, and patient outcome prediction. It is also used to see if a customer will churn or not. Also used for identifying if patient is at developing a particular disease or not. 5  
  
**Random Forest:**

Description: Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. Each decision tree is constructed using a random subset of the data and features. The final prediction is obtained by aggregating the predictions of individual trees (e.g., averaging or voting). Random Forests can be used for classification tasks like predicting whether a diabetic patient will be readmitted to the hospital.4

Advantages:

• High Accuracy: Random Forest tends to produce accurate predictions because it reduces overfitting and variance by averaging the results of multiple trees.

• Handles Mixed Data: It can handle a mix of categorical and numerical features, making it suitable for healthcare datasets that often contain diverse types of information.

• Feature Importance: Random Forest can provide information about the importance of each feature, helping you identify the most influential factors for readmission.4

Use Cases:

• Random Forest is widely used in healthcare for tasks like disease prediction, readmission risk assessment, and patient outcome prediction.

• It's effective in identifying patients who are at a higher risk of readmission, allowing healthcare providers to allocate resources efficiently and provide timely interventions.4

**Random Forest and Decision tree: When to Use Them 3**

• Where goal is to find meaningful subgroups of sample related optimally to a dependent variable - All the given variables affect the readmission of diabetic patients within 30 days. So, both can be of good use.

• When many types of input variables are present: Binary, categorical, ordinal, continuous - As can be seen from the data set given, there are many categorical and binary variables. We can use One hot encoding for binary variables can convert categorical variables into dummy variables. The next step would be to normalize or scale the data and select the important features.

• When large data sets are available to cover the space - We have ‘Diabetic\_data.csv’ (18.2MB) with 101,766 data points. 1 That is a pretty large data set to cover the whole space.

• Where relationships are not linear - In this scenario, the relationships are not linear.

• When algorithms can be run multiple times on subsets of data to obtain robust results

The last and not-so-appropriate analytics technique for predicting hospital readmission for diabetic patients within 30 days would be **Binary Logistic Regression.**

Binary logistic regression: In this approach, the response or dependent variable has only two possible outcomes (e.g. 0 or 1). We can use it in our analytics project to predict if a diabetic patient will be readmitted within 30 days or not. Within logistic regression, this is the most commonly used approach, and more generally, it is one of the most common classifiers for binary classification.2

**Binary Logistic Regression: When to use it?** [MAP F23 Session 5 - Making choices around analytics.pdf]

• When you have a fixed number of categories (2 or more) - Two categories that we will be focusing on are the diabetic patients readmitted to hospital within 30 days and ones who are not.

• When input variable are numeric (or mostly numeric) - Most of the variables in our data set are numeric and categorical values (which can be converted to dummy variables) and then normalized or scaled.

• When you want your results to be interpretable - We want to interpret whether the diabetic patient will be readmitted within 30 days or not based on the data that we have from ‘Diabetic\_data.csv’ (18.2MB) with 101,766 data points. 1

• When you want to understand the relative impact of features - To get the most important features from the multiple features we have, we can use Principal component analysis or Random first for feature selection.

Advantages:

• Interpretability: Logistic regression provides easily interpretable results. You can directly interpret the coefficients of the input features to understand their impact on the likelihood of readmission.

• Low Complexity: It's a relatively simple and computationally efficient algorithm, making it a good choice for initial analysis and when you have a limited dataset.

• Feature Importance: Logistic regression can highlight which features are most influential in predicting readmission by using Principal component analysis. 2

Use Cases:

• Logistic regression is commonly used in healthcare for risk prediction, such as predicting readmission rates, disease outcomes, or the likelihood of complications.

• It can help healthcare providers identify patients at higher risk of readmission, allowing for proactive interventions and better resource allocation. 2

**3. What techniques satisfy the decision and process constraints (accuracy, timeliness, etc.)?**

For our project, the decision is whether a diabetic patient should be discharged from the hospital. This decision would be made daily by the hospital workers.

The process constraints that need to be considered:

* The result should be available in human response times, ideally less than 2 seconds, not more than a minute. Health workers need the answers to the decisions when treating patients; thus, a batch process wouldn’t give timely support. However, real-time results in milliseconds are not required.
* We have a fixed number of categories for the decision: two categories, to discharge or not to discharge an individual patient from the hospital, or the same as a high probability of being readmitted in less than 30 days vs. a low probability of being readmitted in less than 30 days.
* Accuracy is important because the health system wants to minimize False Negatives of being readmitted because a patient who is discharged from the hospital because it was misclassified could generate significant negative consequences for the patient, their families, and the health system. Then, the objective is to maximize the accuracy while providing a suitable solution for the health works operational needs
* The input variables are mostly categorical and numeric: 13 variables are integer type, 34 are categorical types, and 3 are boolean type columns. Examples of categorical variables:
  + Race: Caucasian, African-American, “?”, Other, Asian, Hispanic
  + Gender: Male, Female
  + Age: [0-10), [10-20), [20-30), [30-40), [40-50), [50-60), [60-70), [70-80), [80-90), [90-100)
  + Admission type: Emergency, Urgent, Elective, Newborn, Not Available, NULL, Trauma Center, Not Mapped
  + Metformin: “Not”, “Steady”
* The data exploration exercise from the previous homework showed that the numerical variables follow various distribution functions. For example, the patient time in the hospital and the number of medications variables have a right-skewed distribution, while the number of lab procedures is more symmetrical and closer to a normal distribution.
* The available data have many negative and positive labeled examples. The readmitted dependent variable is present in every data point with values: “<30”, ”>30”, “NO”.
* From the health workers' perspective, techniques that can explain the decision rather than “black box” techniques are preferred. In this case, they prefer a technique that can be interpretable and where they can understand the relative impact of the features

Given these conditions, we can assess if the common techniques used for classification problems could be used for this specific case:

* Binary logistic regression as classification doesn’t completely satisfy the problem because most of the variables are categorical
  + ✔️Works well to predict the probability that the observation matches a category. In this case, the category is being readmitted
  + ❌The input variables are numeric or mostly numerical
  + ✔️Allows to understand the relative impact of features
* Decision trees satisfy the constraints of the problem. However, this technique doesn’t have the best accuracy of the viable alternatives
  + ✔️used to find meaningful subgroups of sample related optimally to a dependent variable – readmitted is the dependent variable
  + ✔️When many types of input variables are present – we have many numeric and a few categorical
  + ✔️When large data sets are available to cover the space – more than 100K data points
  + ✔️Where relationships are not linear - categorical variables are not linear
  + ✔️When algorithms can be run multiple times on subsets of data to obtain robust results - a data snapshot is available and is adequate for multiple runs
  + ✔️Easy to interpret - the resulting three has a clear set of factors
  + ❌Accuracy depends on the training data and can be less accurate than other techniques
  + ✔️Processing times suitable for real time applications
* Random forest produces good results with better accuracy, but it’s difficult to interpret
  + ✔️Used to find meaningful subgroups of sample related optimally to a dependent variable – readmitted is the dependent variable
  + ✔️When many types of input variables are present – We have many numeric and a few categorical
  + ✔️When large data sets are available to cover the space – more than 100K data points
  + ❌Difficult to interpret as the combination of random trees usually produces a complex logic
  + ✔️Processing times aren’t suitable for real-time applications, but they can provide classification predictions in human-readable time

With these problem constraints, the techniques that better satisfy the scenario are decision trees and Random forests. Decision trees can have lower accuracy and are more sensible to the training data, but it’s simpler and easier to understand. Random forests use bootstrap aggregating to improve accuracy, but the resultant model is more difficult to interpret.

We recommend using the decision trees technique for predicting hospital readmission for diabetic patients within 30 days, as in this case, it would have better acceptance and adoption by the medical personnel because it's easy to interpret.

**4. What information needs to be communicated to stakeholders to take the necessary action(s)? The answer should describe the result of the analytics that the decision-maker uses to make an improved decision, and perhaps information relevant to other stakeholders as well (say, cost or accuracy of the decision)**

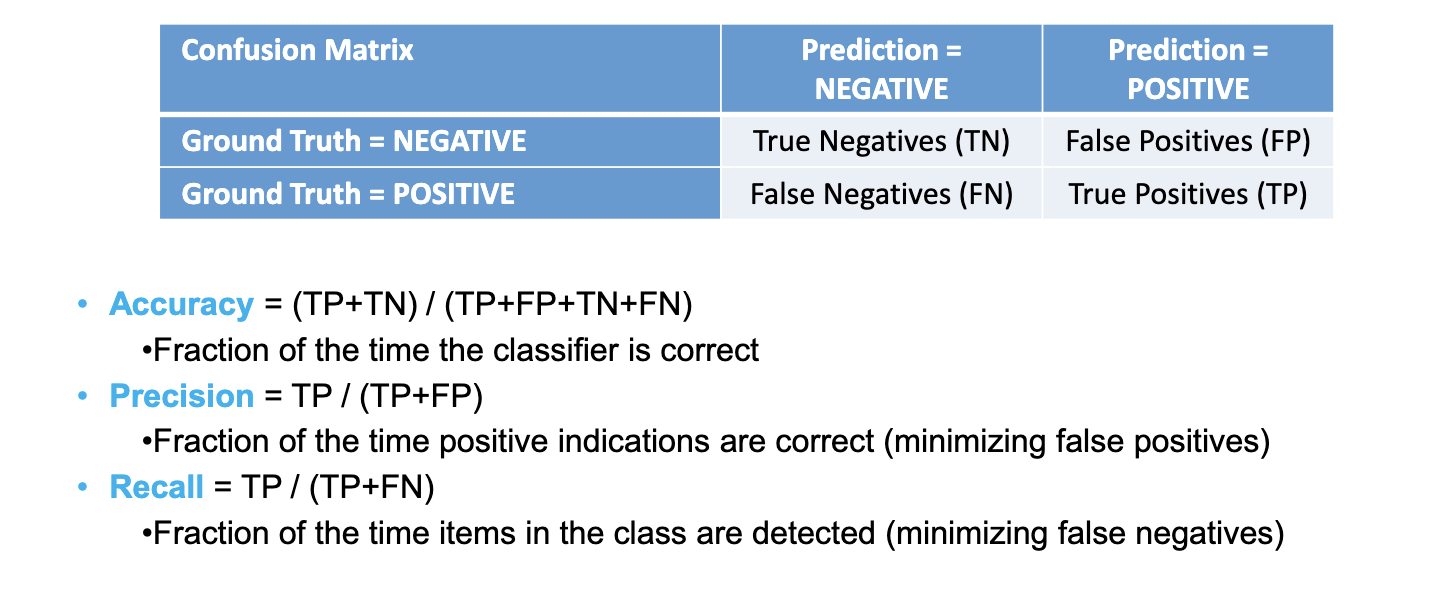
For stakeholders, the first step is explaining and visualizing data in a simple and understandable manner. Some examples of visualizations will be provided in the next question.

Next, the reasoning behind choosing the model should be explained (refer to questions 2 and 3) as well as the mechanisms behind the models in understandable terms even for a non-technical audience. For example, logistic regression explained in simple terms is statistical model that outputs the probability of an event based on relevant variables/features. In our case, we want to predict if diabetic patient should be readmitted to the hospital within 30 days based on features such as the number of medications, number of days between admission and discharge, and number of lab tests performed during the encounter. For our model, the probability the diabetic patient is readmitted and the probability the diabetic patient is not readmitted sums up to 1. If our model probability is greater than .5 for the patient being readmitted, we can classify the patient into this category. In simple terms, a decision tree is a set of decision rules in a tree structure that helps predict an outcome variable (whether the patient should be readmitted) based on relevant categorical/numerical variables. A random forest is a collection of decision trees combined into one model.

*We will then evaluate our classification performance (accuracy, precision, and recall) and present it to stakeholders.*

In our case:

* Ground Truth = Negative: Diabetic patient was not readmitted within 30 days
* Ground Truth = Positive: Diabetic patient was readmitted within 30 days
* Prediction = Negative: Model gave a higher probability for the patient to not be readmitted within 30 days
* Prediction = Positive: Model gave a higher probability for the patient to be readmitted within 30 days

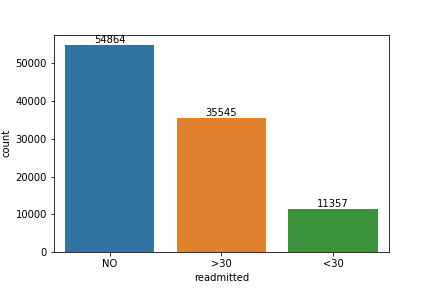


*From Lecture 5*

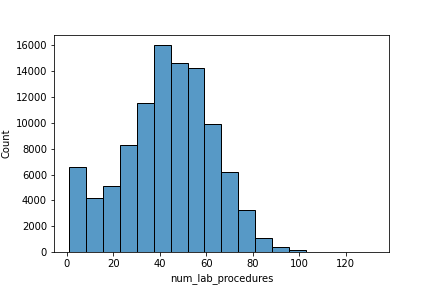
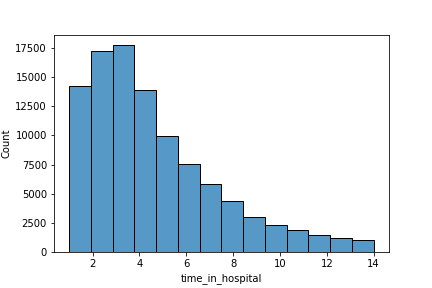
Moreover, from our model, we can obtain feature weights. These feature weights help assess the significance of each feature in predicting the outcome variable. If the value of the feature weight is larger, this indicates that the feature is more important in predicting the outcome variable in our model. A positive weight indicates that as the value of the feature increases, the probability of our target variable being 1 (patient being readmitted) increases. A negative weight indicates that the probability of our target variable being 0 (patient not being readmitted) increases.

**5. Which data visualization techniques best communicate the necessary information for each stakeholder? Full answers should be specific visualization techniques, like bar charts, or Sankey diagrams. The answers include an example applied to your project, even in prototype form if actual data is not available.**

*There are many different visualizations we could use, such as bar charts for comparison of values and histograms for distribution of values.*

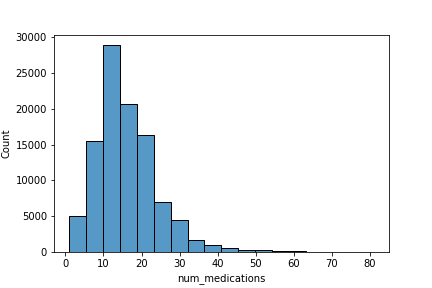


**Figure 1: Counts for Readmission**

**

**Figure 2: Histogram of Patient Figure 3: Histogram of Number**

**Time in Hospital (Days) of Lab Procedures**



**Figure 4: Histogram of Number**

**of Medications**

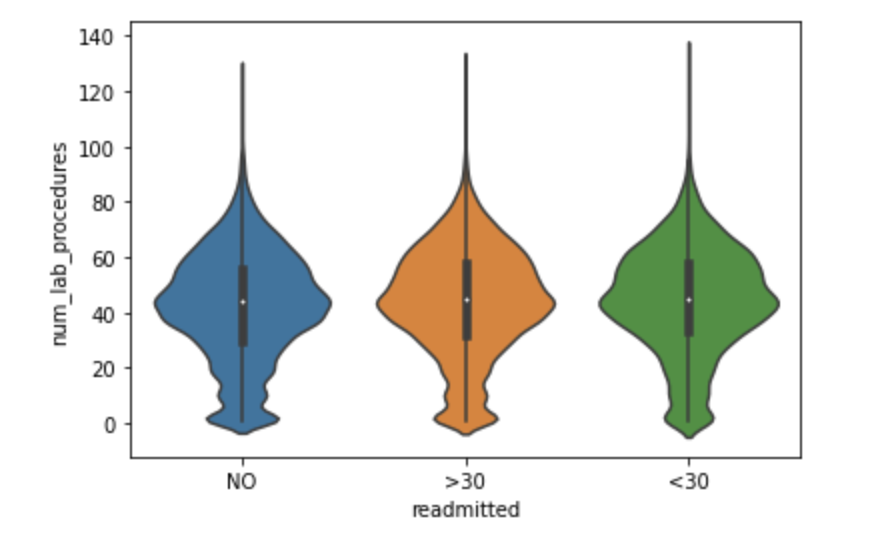
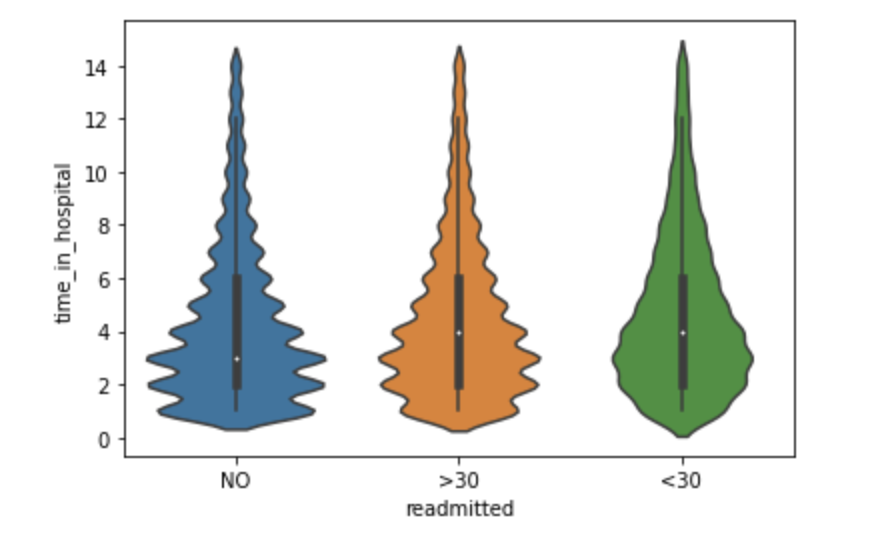
Figure 1 displays the counts for readmission. The majority of patients had readmission marked as “No”, meaning that there was no record of readmission. There was a significantly higher number of patients (~24,000) readmitted in more than 30 days than less than 30 days.

FIgure 2 displays a histogram of hospital time, the number of days between admission and discharge. This histogram is skewed right with a peak around 3-4, meaning that most patients stay in the hospital around 3-4 days.

Figure 3 displays a histogram of the number of lab procedures. The histogram does not display a normal distribution as there’s a significant count approximately below 10 lab procedures as seen in the left-most bar. The peak of the histogram is around 40, meaning that most patients receive around 40 lab procedures.

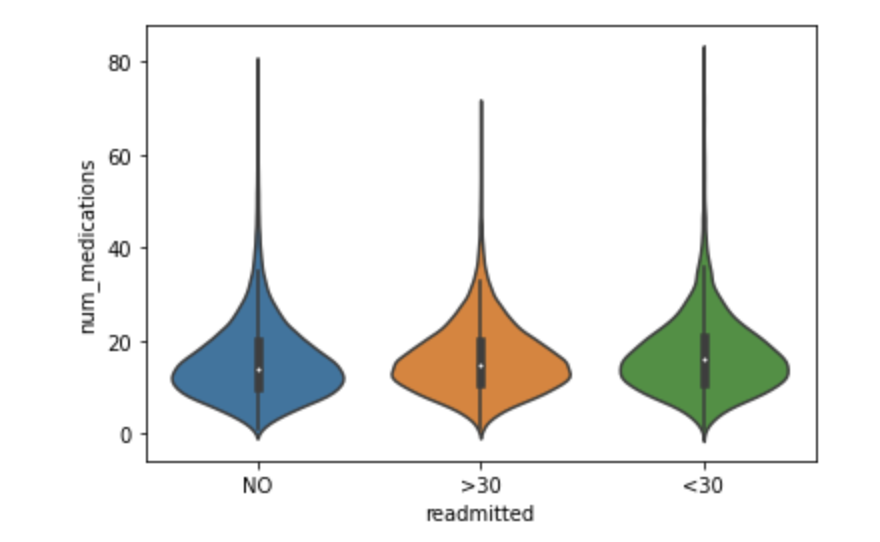
Figure 4 displays a histogram of the number of medications. The histogram is not heavily skewed. The peak is at around 10-15, meaning that most patients receive about 10-15 medications.

*To visualize the relationship between some numerical variables and our categorical outcome variable, we make use of violin plots.*



**Figure 5: Violin Plot of Time in Hospital Figure 6: Violin Plot of Number of Lab**

**by Readmission Category Procedures by Readmission Category**

****

**Figure 7: Violin Plot of Number of Medications**

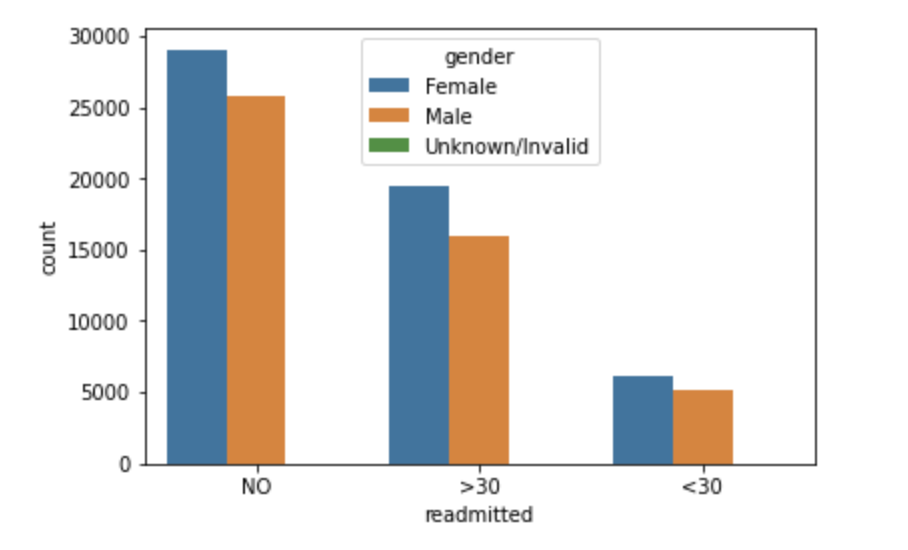
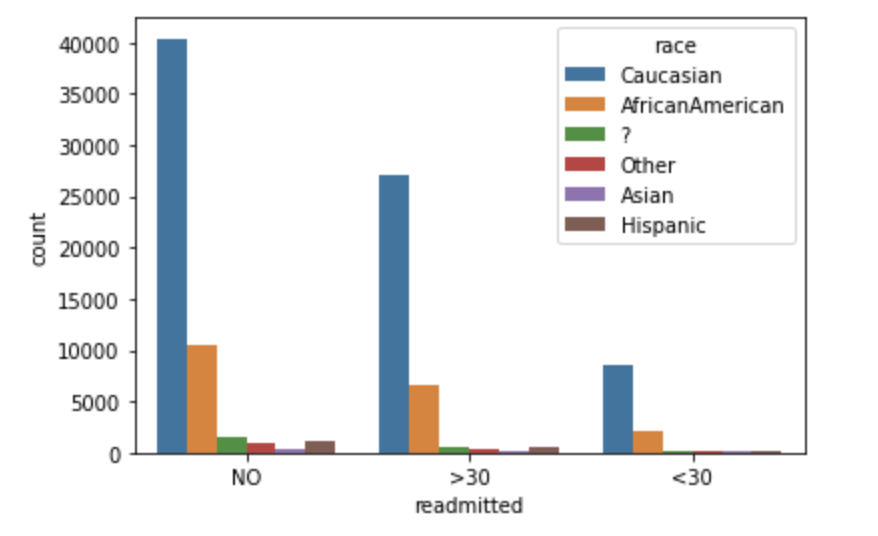
**by Readmission Category**

Figure 5 displays the distribution of time in hospital by readmission category. The shape of the distribution for time in hospital is notably different with less curves for those readmitted in less than 30 days.

Figure 6 displays the distribution of the number of lab procedures by readmission category. For those that were admitted <30 days, there are less curves in the distribution around the lower end of number of lab procedures.

Figure 7 displays the distribution of the number of medications by readmission category. For those that were admitted >30 days, the highest number of medications was slightly lower than that of the different categories.

*To visualize the relationship between some categorical variables and our categorical outcome variable, we make use of stacked bar plots.*



**Figure 8: Stacked Bar Plot of Race Figure 9: Stacked Bar Plot of Gender**

**by Readmission Category by Readmission Category**

Figure 8 displays a stacked bar plot of race by readmission category. Majority of patients were Caucasian across all the categories.

Figure 9 displays a stacked bar plot of gender by readmission category. There is less of a difference in gender counts for patients readmitted in <30 days.

**References**

1. Clore, John, Jon DeShazo, and Beata Strack. 2014. “Diabetes 130-US hospitals for years 1999-2008.” UCI Machine Learning Repository. <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>.
2. “What is Logistic regression?” n.d. IBM. Accessed October 3, 2023. <https://www.ibm.com/topics/logistic-regression>.
3. MAP F23 Session 5 - Making choices around analytics.pdf
4. “What is Random Forest?” n.d. IBM. Accessed October 3, 2023. <https://www.ibm.com/topics/random-forest>.
5. Hogarty, Steve. 2022. “Decision trees: Definition, analysis, and examples.” WeWork. <https://www.wework.com/ideas/professional-development/business-solutions/decision-trees-definition-analysis-and-examples>.
6. “1.10. Decision Trees — scikit-learn 1.3.1 documentation.” n.d. Scikit-learn. Accessed October 4, 2023. <https://scikit-learn.org/stable/modules/tree.html>.